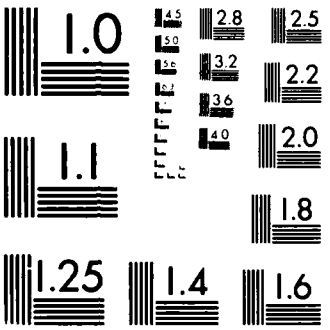
AD-A149 565 RESEARCH ON KNOWLEDGE DELIVERY(U) UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST W C MANN DEC 84 ISI/SR-84-148 F49620-79-C-0181 1/1

UNCLASSIFIED F/G 9/2 NL



ISI Special Report
ISI/SR-84-148
December 1984

William C. Mann

Research on Knowledge Delivery

Contract Final Report of Research in the period August 15, 1979 to August 14, 1984.

INFORMATION SCIENCES INSTITUTE

4676 Admiralty Way • Marina del Rey, California 90291

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ISI/SR-84-148	2. GOVT ACCESSION NO. AD-A149 565	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* Research on Knowledge Delivery		5. TYPE OF REPORT & PERIOD COVERED Final Report August 15, 1979–August 14, 1984
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) William C. Mann		8. CONTRACT OR GRANT NUMBER(s) F49620-79-C-0181
9. PERFORMING ORGANIZATION NAME AND ADDRESS USC/Information Sciences Institute 4676 Admiralty Way Marina del Rey, CA 90292-6695		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research Building 410, Bolling Air Force Base Washington, DC 20332		12. REPORT DATE December 1984
		13. NUMBER OF PAGES 15
14. MONITORING AGENCY NAME & ADDRESS(*if different from Controlling Office*)		15. SECURITY CLASS. (*of this report*) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

This document is approved for public release; distribution is unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

.....

18. SUPPLEMENTARY NOTES

.....

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

artificial intelligence, computational linguistics, discourse, grammar, Inquiry Semantics, natural language, Nigel, Penman, Relational Propositions, rhetoric, Rhetorical Structure Theory, semantics, Systemic Linguistics, text generation, text planning, text structure

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

The general practical goal of this work is to enable computers to express information in appropriate, easily read, multiparagraph English text. This report describes the progress of the Knowledge Delivery project toward that goal, focusing on three technical areas: computer grammars of English for text generation, control structures for text generation, and the description and construction of discourse structures. The project's accomplishments in these three areas are expressed in the Nigel grammar, the Penman system, and the Rhetorical Structure Theory, respectively.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

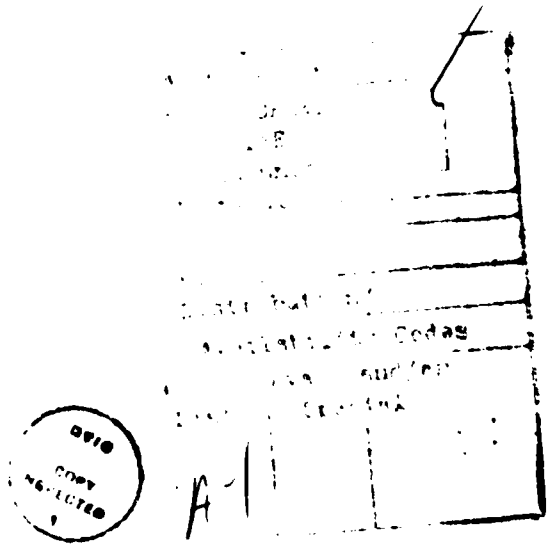
ISI Special Report
ISI/SR-84-148
December 1984

University of Southern California

William C. Mann

Research on Knowledge Delivery

Contract Final Report of Research in the period August 15, 1979 to August 14, 1984.



INFORMATION SCIENCES INSTITUTE

213/822-1511

4676 Admiralty Way/Marina del Rey/California 90292-6695

This research was supported by the Air Force Office of Scientific Research Contract No. F49620-79-C-0181. The views and conclusions contained in this paper are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research of the U.S. Government. This report is written according to a contractually prescribed criterion.

Research on Knowledge Delivery

William C. Mann
USC Information Sciences Institute
December 1984

1 Objectives

The general practical goal of this work is to enable computers to express information in appropriate, easily read, multiparagraph English text.[1] This goal has never been accomplished with any generality. Computers have generated text only for very narrow domains of subject matter, and always with many handcrafted computational parts.[2]

In order to reach this practical goal, several difficult theoretical problems must be solved:

1. The vocabulary must be characterized, and its relation to available knowledge in the computer must be specified in detail.

2. The grammar of English must be characterized, especially the close alternatives among grammatical constructions and the relationships between particular grammatical constructions and particular effects to be produced by the generated language.

3. The reasons for selecting particular grammatical constructions must be characterized.

4. Reasons for expressing things in a particular order must be identified.

5. Reasons for grouping ideas together in text, and the specific effects of doing so, must be identified.

6. The reasons for choosing to express or withhold particular information must be identified. (This has been a dominant problem in past work on text generation; computers carry so much information on particular topics that very strong selectivity is needed.)

---

[1] The need exists just as much for languages other than English, but English is in many ways the opportunistic target.

[2] There are long computer-generated texts created by a fill-in-the-blanks method, but there the real author is a person.

7. People (and eventually machines) produce text for definite reasons. An appropriate notation for text goals is needed.

8. The processes that produce text are extremely complex. They require appropriately complex and flexible control methods.

The fields that contribute the most to this investigation are computer science and linguistics. The linguistic knowledge of the ways language goes together and produces effects is particularly vital.

Nearly all linguistic investigations are done in a way that characterizes actual text in terms of particular abstractions. They describe text, but do not attempt to say how text might be constructed. Because description is a kind of abstraction, it loses information. Many of the regularities and patterns of good text are not accounted for by linguistic description, yet they are required when good text is constructed. They must therefore be accounted for in any comprehensive text generation process. These undescribed gaps are a significant source of difficulty in text generation research. Their presence helps explain why progress of the field has been so slow, and why so much precomputational theoretical work must be done in creating an effective text generator.

The complexity of the problem makes it important to consider a second kind of goal: Research results must somehow be carried forward from one investigation to another. Journal articles and other publications are necessary, but inadequate, since so much of the created knowledge resides in the details of particular computer programs. The programs themselves must be passed around the research community in a usable form. For the community as a whole, this is a largely unsolved problem. For this project, we have made some progress by creating an elaborate portable research grammar of English.

To reach the more general goals of text generation research, we have focused on more specific project goals:

1. Designing an overall program structure in which grammar, programmed knowledge of discourse, and other components can be appropriately related and controlled.

2. Creating a computational grammar of English for text generation that is transferable from one research group and purpose to another.

3. Creating an explicit theory of discourse structure, suitable as a basis for computer program implementation.

2 Accomplishments and Progress

The accomplishments of the project can be divided into three areas, which correspond to the goals above and are described in the three subsections immediately below.[3]

2.1 Penman: an Evolving Design

The project's text generation system is called Penman. Parts of Penman are programmed, and other parts are at various levels of specification as program designs. One of Penman's modules is Nigel, the grammar. This discussion of Penman concerns its overall control structure; the next section describes Nigel.

The most common sort of generator design moves the information through a pipeline. It starts with an initiating request or goal, then proceeds to a search for things to say, then an organizing phase, and finally production as a set of sentences. This design has several deficiencies, which are corrected in the Penman design.

1. One deficiency concerns the relationship between search and presentation. If the search for things to say is completed before the organization phase (as in the pipeline design), search cannot be used to solve problems of text organization such as the need for background information, the need for evidence for doubtful points, or the need to concede away some apparent refutation. Early versions of Penman did all content discovery before starting to organize the material, incorporating this defect.

---

[3]Publication citations are all deferred to the publications section, Section 3.

In the current Penman design,[4] discovery of things to say is subordinated to text planning. As part of the organization process there are specific limited searches for particular kinds of information. This is more efficient and more effective than earlier designs.

2. A second control design problem concerns the degree to which the effects of particular decisions can be anticipated. As in every complex system, some consequences of decisions are effectively unpredictable. In the pipeline design, decisions in early modules are never assessed or corrected. The text produced by pipeline text generation systems thus contains characteristic defects.

   In Penman, a specific module is devoted to text improvement. It assesses the text, checking for recognized patterns of defect. It then selects a high-priority defect to correct, modifies the text plan accordingly, and regenerates the changed portions. The text is complete only when this "hindsight module" cannot produce improvements. This approach to control overcomes the problem of weak anticipation of the consequences of decisions.

2.2 The Nigel Text Generation Grammar

When this project was begun, there was no single preeminent variety of grammar for text generation. There are many grammatical formalisms, but only a few relate grammatical structure to the effects the language produces. The Systemic Functional Grammar framework of Systemic Linguistics was chosen for this work. Fortunately, shortly after the Systemic framework was chosen, its founder (Michael Halliday) joined the project as a consultant.

At the beginning of this research, Systemic Linguistics had described many portions of English in separate fragments of grammar, but (aside from Hudson's work [Hudson 71],

---

[4] The most recent public design report reflects the older approach; the next one will correct it.

which we did not follow in detail) there was no comprehensive, consistent treatment in a single notation. The grammar we have built, called Nigel, is now the most comprehensive grammar of English for text generation, necessarily in a single notation since it is programmed.

Before Nigel was built, systemicists had not yet agreed on the details of how grammatical structures are constructed; Nigel has a more explicit and formal notation for structure building than any predecessor.

Beyond these refinements of the grammatical notation, the work on Nigel has extended the Systemic framework so that its semantics is much more explicit than before. Systemic grammar is organized around choice among alternatives. In the development of a systemic grammar, sets of closely related alternatives are identified and the patterns of alternatives are described in systemic notation. However, describing the alternatives does not include describing the circumstances under which each particular alternative is chosen. Although those circumstances are often described informally, the grammar's relation to situations of language use is not made very explicit in that way.

To make the choice processes explicit and programmable, as part of the development of Nigel, a new formal notation for choice among alternatives has been created. This notation and the choice procedures expressed in it are called *Inquiry Semantics*. In addition to facilitating programming, Inquiry Semantics is an extension of the Systemic Linguistics framework.

In a major test of the Inquiry Semantics framework, choice procedures have been created for all of the systems of Nigel. The result of this extensive effort is a major explication of the uses of the grammatical alternatives of English.

With this approach to grammar, the generated language can be very well adapted to running text. Other approaches concentrate far more on isolated sentences, but Systemic

Linguistics has built grammars which have extensive provisions for making text flow smoothly. This will be particularly helpful in generating multiparagraph text.

Among computational grammars of natural languages such as English, Nigel is one of the most comprehensive. We therefore expect that, as we disseminate it to appropriate research centers and developers, a wide range of technical and practical goals will be facilitated.

2.3 Rhetorical Structure Theory

It is generally recognized that computer-generated text must be *planned.* The knowledge of text planning is currently rather primitive, partly because related academic topic areas (such as discourse analysis) have focused exclusively on analysis and description, not on synthesis.

For a computer to be able to plan text, there must be an underlying theory of text which is the basis for the program. At the beginning of this research, no suitable theories existed. What was known was too vague and analytically oriented to use as the basis for program design.

We have created a new body of theory, called ***Rhetorical Structure Theory*** (RST), which represents the structure of written monologues such as reports, magazine articles, instructions, and other expository kinds of text. The theory is currently descriptive rather than constructive, i.e., it does not yet show how to do all the steps of text planning. It does specify what a text plan must contain, and informal experiments have shown that it is useful as a basis for text creation by people. On a large scale, it has been used to plan technical reports and conference papers. On a smaller scale, it has been used to show how particular existing texts could have been created by a programmed generator. These informal experiments demonstrate a level of definiteness and planfulness beyond what the published art had previously achieved, and they show that RST is suitable as a text planning design base.

2.4 Summary of Progress

The project has produced several of the major technological prerequisites to creating a general-purpose, programmed generator of English text. On the basis of these accomplishments, several new projects are now working on various aspects of the problem. Added work is needed in the areas of lexical selection, noun phrase planning, sentence planning, interfacing Nigel to various programmed knowledge notations, text planning, text brevity techniques, and appropriate representation of the static and dynamic knowledge of the reader.

In the follow-on projects already under way, we will distribute the Nigel grammar to other research organizations and will build several text generators for specific purposes.

3 Publications and Planned Publications

The publications and planned publications of the project are described below, and full references are given in the bibliography.

1. [Matthiessen 81] -- *A Grammar and Lexicon for a Text-production System* -- describes how Nigel can be interfaced to a lexicon and associated knowledge base.

2. [Mann 81a] -- *Text Generation: The State of the Art and the Literature* -- the first comprehensive survey of text generation technology, with the first comprehensive bibliography and an account of projects then in progress. Coauthors are Madeline Bates, Barbara Grosz, David D. McDonald, Kathleen R. McKeown, and William Swartout of BBN, SRI, UMass Amherst, University of Pennsylvania, and ISI. Published in *American Journal of Computational Linguistics* under the title "Text Generation."

3. [Mann 81b] -- *Two Discourse Generators* -- compares two text generators that were created prior to this research, Davey's Proteus and Mann and Moore's KDS.

4. [Mann 82] -- *Anatomy of a Systemic Choice* -- describes how the Systemic notational framework can be extended to express more of the semantics of a grammar, in particular, supplementing the expression of grammatical alternatives with a formal expression of the circumstances under which each alternative is chosen. Appears in highly abbreviated form in the Coling 1982 Proceedings.

5. [Mann 83a] -- *An Overview of the Penman Text Generation System* -- describes Penman's general design and identifies operations that are required to support general text generation.

6. [Mann 83b] -- *Systemic Encounters with Computation* -- a historical survey of all of the computational work done in the Systemic framework.

7. [Mann 83c] -- *Inquiry Semantics: A Functional Semantics of Natural Language Grammar* -- describes the interactions between Nigel and its informational environment, and shows how these interactions determine what Nigel says.

8. [Mann 83d] -- *A Linguistic Overview of the Nigel Text Generation Grammar* -- describes Nigel for a linguistic audience.

9. [Mann 83e] -- *An Introduction to the Nigel Text Generation Grammar* -- basic orientation to Nigel's capabilities, notation, and conte

10. [Mann & Thompson 83] -- *Relational Propositions in Discourse* -- describes varieties of assertions that arise in running discou ut not in isolated sentences. Establishes several categories of assertions that can be made from discourse structure alone, without explicit signals.

11. [Matthiessen 83a] -- *Choosing Tense in English* -- a Systemic treatment of tense in English, based on Nigel.

12. [Matthiessen 83b] -- *Choosing Primary Tense in English* -- a journal article containing a Systemic treatment of English primary tense, based on [Matthiessen 83a].

13. [Matthiessen 83c] -- *Choosing Tense in English* -- a full description of Nigel's treatment of English tense, expanding on [Matthiessen 83a] to include justifications, additional detail, and comparison with other approaches.

14. [Matthiessen 83d] -- *The Systemic Framework in Text Generation: Nigel* -- a linguistically oriented description of how the general desiderata of the Systemic framework interact with the text generation task, taking Nigel as the central example.

15. [Mann 84] -- *Discourse Structures for Text Generation* -- describes Rhetorical Structure Theory as a basis for text planning, and compares it with prior computational text planning work.

16. [Matthiessen 84] -- *What's in Nigel: 1* -- a brief introduction to the content of Nigel for functional linguists.

17. [Mann & Matthiessen 84] -- *The Realization Operators of the Nigel Grammar* -- a brief introduction to Nigel's methods of structure building for functional linguists.

In addition, project members have contributed to conference presentations, invited lectures, workshops, and other presentations for which there were no written proceedings. All of the ISI reports in the bibliography have been submitted to the Defense Technical Information Center, and most are still available from ISI.

Planned publications include the following:

1. a paper on the Relational Transitivity portion of the Nigel grammar. This paper is in a category of papers which have been specifically requested from us by a leading journal. We expect to publish in that journal.

2. a paper on the relationship between Rhetorical Structure Theory and covert communication (unsignalled assertions which do not correspond to clauses).

3. a report on English tense. This is a major expansion of Christian Matthiessen's thesis [Matthiessen 83a], including new material and reflecting in more detail the treatment of tense in Nigel.

4. a brief introduction to Inquiry Semantics for functional linguists.

It is likely that (beyond the expected publications above) two books will also be written based on this work, one on the grammar and one on discourse, but these plans are not definite enough to describe. Prospective tables of contents for each have been developed.

4 Professional Personnel and Awarded Degrees

The following professional personnel have contributed to this project:

1. Michael Fehling
2. Barbara Fox
3. Yasutomo Fukumochi
4. Michael Halliday
5. Steve Klein
6. William Mann
7. Christian Matthiessen
8. James Moore
9. Sandra Thompson

In addition, we have had many professional visitors, including several who have contributed from a week to over a month of their time to the project as unpaid consultants. Their contributions have been substantial. These visitors include the following:

1. Peter Fries
2. Ruqaiya Hasan
3. James Martin
4. Robert Spence
5. R. MacMillan Thompson
6. David Weber

Degrees have been awarded to two of the project staff. In both cases, technology developed on the project was a part of the work done to qualify them for their degrees.

Barbara Fox
PhD in Linguistics
University of California at Los Angeles (UCLA)
1984

Christian Matthiessen
Master of Arts in Linguistics
University of California at Los Angeles (UCLA)
1983

Christian Matthiessen's thesis, entitled *Choosing Tense in English*, was based on the Nigel grammar.

References

[Hudson 71] Hudson, R. A., *North-Holland Linguistic Series.* Volume 4: *English Complex Sentences*, North-Holland, London and Amsterdam, 1971.

[Mann 81a] Mann, W. C., *et al.*, *Text Generation: The State of the Art and the Literature*, USC/Information Sciences Institute, RR-81-101, December 1981. Appeared as *Text Generation* in April-June 1982 *American Journal of Computational Linguistics.*

[Mann 81b] Mann, W. C., "Two discourse generators," in *The Nineteenth Annual Meeting of the Association for Computational Linguistics*, Sperry Univac, 1981.

[Mann 82] Mann, W. C., *The Anatomy of a Systemic Choice*, USC/Information Sciences Institute, RR-82-104, October 1982. To appear in *Discourse Processes.*

[Mann 83a] Mann, W. C., "An overview of the Penman text generation system," in *Proceedings of the National Conference on Artificial Intelligence*, pp. 261-265, AAAI, August 1983. Also appears as USC/Information Sciences Institute, RR-83-114.

[Mann 83b] Mann, W. C., "Systemic encounters with computation," *Network: News, Views and Reviews in Systemic Linguistics and Related Areas*, (5), May 1983, 27-33.

[Mann 83c] Mann, W. C., "Inquiry semantics: A functional semantics of natural language grammar," in *Proceedings of the First Annual Conference*, Association for Computational Linguistics, European Chapter, September 1983.

[Mann 83d] Mann, W. C., "A linguistic overview of the Nigel text generation grammar," in *Proceedings of the Xth International LACUS Forum*, Linguistic Association of Canada and the United States, Quebec City, Quebec, Canada, August 1983.

[Mann 83e] Mann, W. C., "An introduction to the Nigel text generation grammar," in *Nigel: A Systemic Grammar for Text Generation*, USC/Information Sciences Institute, RR-83-105, February 1983. This paper will also appear in a forthcoming volume of the *Advances in Discourse Processes Series*, R. Freedle (ed.): *Systemic Perspectives on Discourse: Selected Theoretical Papers from the 9th International Systemic Workshop*, to be published by Ablex.

[Mann 84] Mann, W. C., *Discourse Structures for Text Generation*, USC/Information Sciences Institute, RR-84-127, February 1984. Also appeared in the proceedings of the 1984 Coling/ACL conference, July 1984.

[Mann & Matthiessen 84] Mann, W. C., and C. M. I. M. Matthiessen, "The realization operators of the Nigel grammar," *Network: News, Views and Reviews in Systemic Linguistics and Related Areas*, (7), 1984, to appear.

[Mann & Thompson 83] Mann, W. C., and S. A. Thompson, *Relational Propositions in Discourse*, USC/Information Sciences Institute, RR-83-115, July 1983. To appear in *Discourse Processes.*

[Matthiessen 81] Matthiessen, C. M. I. M., "A grammar and a lexicon for a text-production system," in *The Nineteenth Annual Meeting of the Association for Computational Linguistics*, Sperry Univac, 1981.

[Matthiessen 83a] Matthiessen, C. M. I. M., Choosing Tense in English, Master's thesis, University of California at Los Angeles, 1983.

[Matthiessen 83b] Matthiessen, C. M. I. M., "Choosing primary tense in English," *Studies in Language* 7, (3), 1983.

[Matthiessen 83c] Matthiessen, C. M. I. M., Choosing Tense in English, 1983.

[Matthiessen 83d] Matthiessen, C. M. I. M., "The systemic framework in text generation: Nigel," in *Nigel: A Systemic Grammar for Text Generation*, USC/Information Sciences Institute, RR-83-105, 1983. This paper will also appear in a forthcoming volume of the *Advances in Discourse Processes Series*, R. Freedle (ed.): *Systemic Perspectives on Discourse: Selected Theoretical Papers from the 9th International Systemic Workshop*, to be published by Ablex.

[Matthiessen 84] Matthiessen, C. M. I. M., "What's in Nigel: 1," *Network: News, Views and Reviews in Systemic Linguistics and Related Areas*, (6), April 1984, 36-44.

END

FILMED

2-85

DTIC